Hourly energy demand and prices: an analysis on risk measures and correlation

Camilo Oberndorfer Mejía Gregorio Pérez Bernal Luisa Toro Villegas

Miguel Valencia Ochoa

May 27, 2021

Contents

Abstract

blah blah lo que vamos a usar blah blah.

The previously mentioned methodologies will be applied to the hourly energy demand generation in Spain from 2015 to 2019. The hope to this paper is to find a way to predict energy prices using correlation techniques and understand how demand affects supply and vicersa.

1 Introduction

code at [lto21]

2 Data description

2.1 Original dataset

This dataset included 35048 data of 20 different variables about electric energy generation, demand, price and forecasts of some of these variables.

It included how much energy was generated by every generation method used in Valencia, Spain, how much demand the energy had and the price of the energy at a given time. Every hour of every day from January 1st of 2015 to December 31 of 2018 has its own data consisting on the previously mentioned variables excluding those hours in which the energy generation, demand or price was not measured.

For more detailed reading about the original dataset, it can be found at:[Jha19]

2.2 Modifications

For each method used in this paper, a subset of the original data was chosen according to each specific needs, mainly due to the need to visualize properly in graphs. These subsets are:

- A subset of the data including only the information of the hours of the last month on the dataset which was called data720 including exactly 720 data registries.
- A subset including the data from the first hour of everyday was included in the construction of the control charts, this data includes 1460 data registries and is called data1460.
- A data subset including the data from the first hour of each month was taken, this data set included 48 data and therefore is called data48.

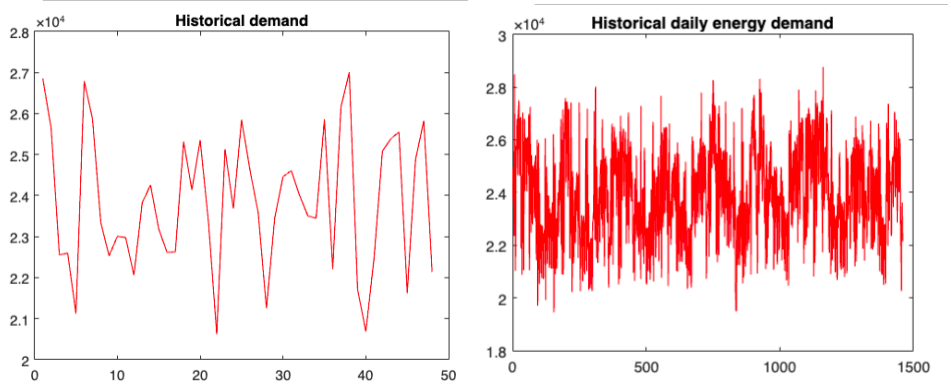
A second dataset that included weather in different regions of Spain during that time. The only data considered was that of Valencia. A problem that was encountered was that several entries of the dataset had categorical values, which means that analysing them was more difficult. This is why methods of Categorical Encoding were used. Basically, the method consists on assigning a numerical value for every non-numerical input. For instance, if the entries regarding climate were: "Sunny", "Rainy", and "Cloudy", each of them would have a numerical value assigned to represent them. It is better to assign them hierarchically, in order for them to have a more realistic way of being analysed.[Yad19]

3 Estimation and results

3.1 Historical data

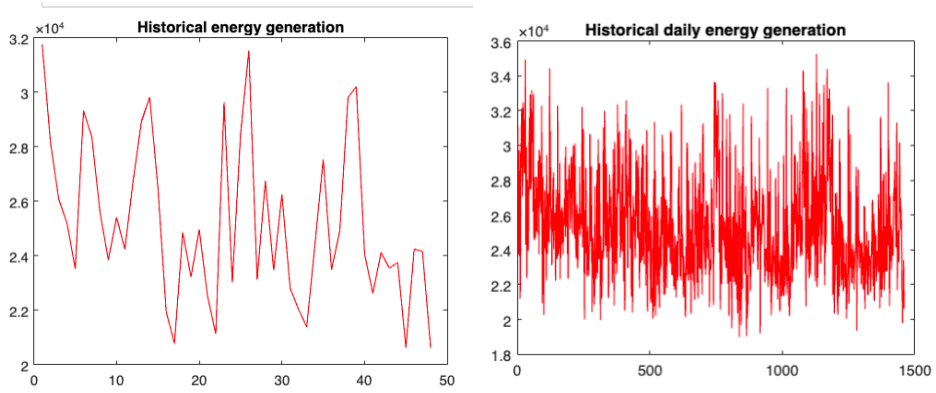
For the historical data, the three main variables of interest (energy demand, energy generation and energy price) were plotted to understand their structure and historical behavior. On the left are the plot for the monthly data and on the right for the daily data of each variable mentioned.

The hourly data's plot is not included because 35000 entries would look too crowded to understand and it provides no extra information. The prediction's data isn't shown either because their behavior is very similar to what they are predicting (specially for demand).

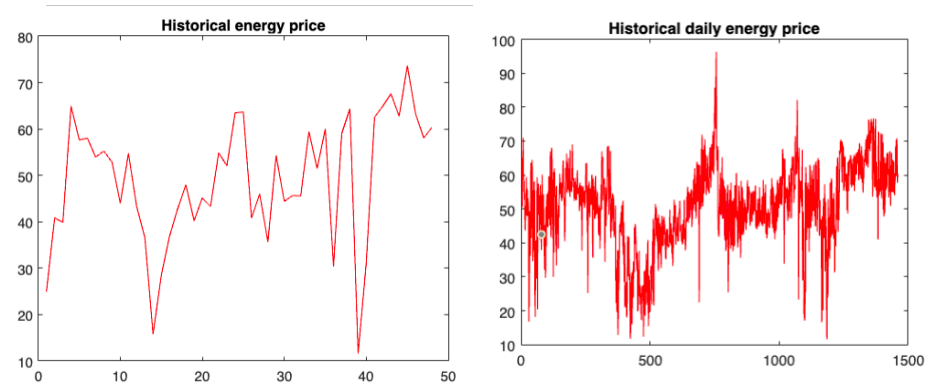


The historical data graph on the left depicts the energy demand of the first hour of every month, as seen, due to the little amount of data, any behaviour or structure can't be differentiated.

However the graph on the right shows the demand included in the dataset data1460 and shows a more clear behaviour were in over all maintains an oscillatory behaviour, this is probably due to weather seasons, making winter and summer the seasons with more electric energy usage, anyways there is still a lot of noise.



There is shown a clear drop in the generation of electrical energy on both graphs but in the graph to the right it has a more clear oscillatory behaviour similar to the one seen in the demand, later its was realized that the most correlated variable to the generation of energy was the forecast of the demand so its predictable for them to have a similar behaviour.

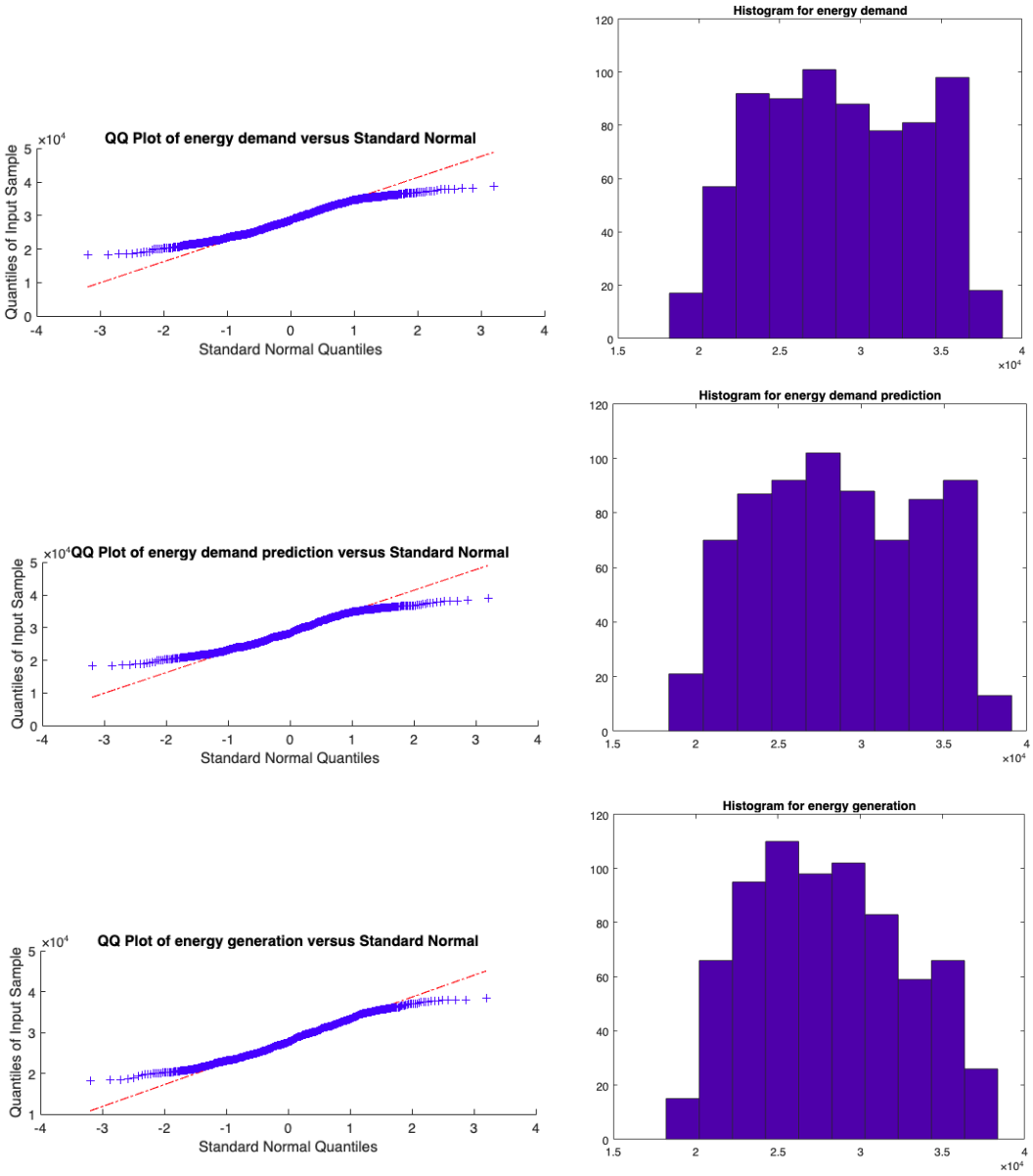


Both graphs show a general slowly rising behaviour on the price of energy with three periods of time were the price changes abruptly, the first drop on the price of energy was probably caused by a previous excess energy generation, this drop itself caused a big demand increase and this later caused the price to go high up again, also a response behaviour is seen as right after the price goes up or down it takes the opposite direction.

3.2 Identifying underlying distribution

To construct a thorough analysis on any given set of data, a crucial first step is to attempt to understand that way the data is distributed. If the data follow a normal distribution, the following analysis could have a smaller margin of error, but it is rare to see that any real life data follows a normal distribution.

The data's energy generation histogram and quantile-quantile plot (qqplot) are graphic tools that help identify normality, which are shown in the following figures:



The previous figures are a characteristic example of under-dispersed data, for example a uniform distribution or a beta distribution with parameters close to (1,1) and the qqplot, which compares the theoretic quantiles and sample quantiles, shows an s-shape, which (light tailed) bleah blah blah.

Hacemos linealizacion? SII!!! hay que hacerlo

3.3 Stability tests

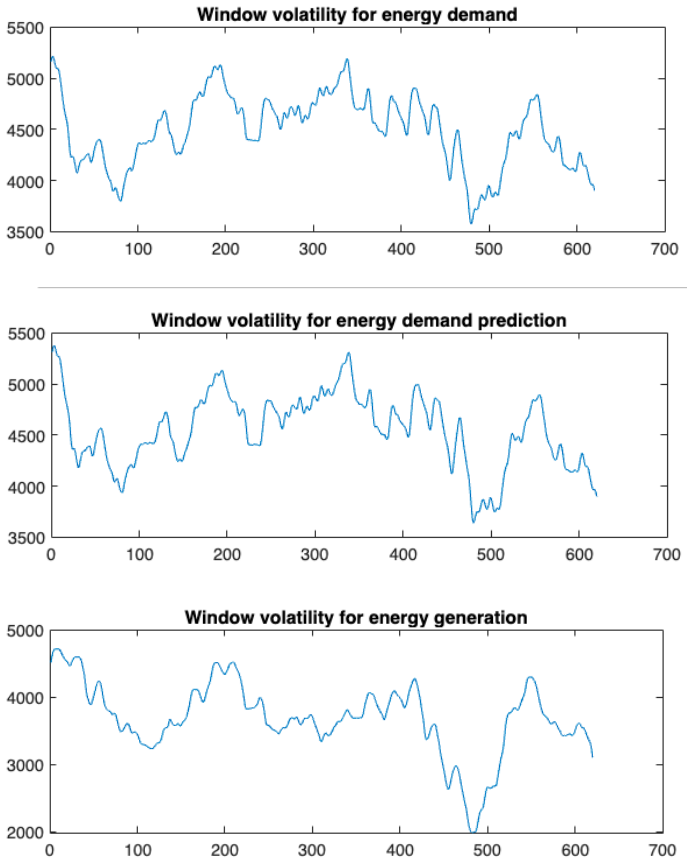
3.3.1 Rolling windows

Rolling windows for volatility is a method that allows to capture the behavior of the variables of interest over time, specially considering that in real-life data parameters such as standard deviation don't stay constant but instead, vary over time. Rolling windows are estimated using the following formula:

$$\sigma^2 = \sum_{i=t-m+1}^{t} \frac{ri^2}{m}, t = m, m+1, ..., n$$

[MB06]

The following graphs show the standard deviation for each of the variables of interest with windows of $m = 100$.

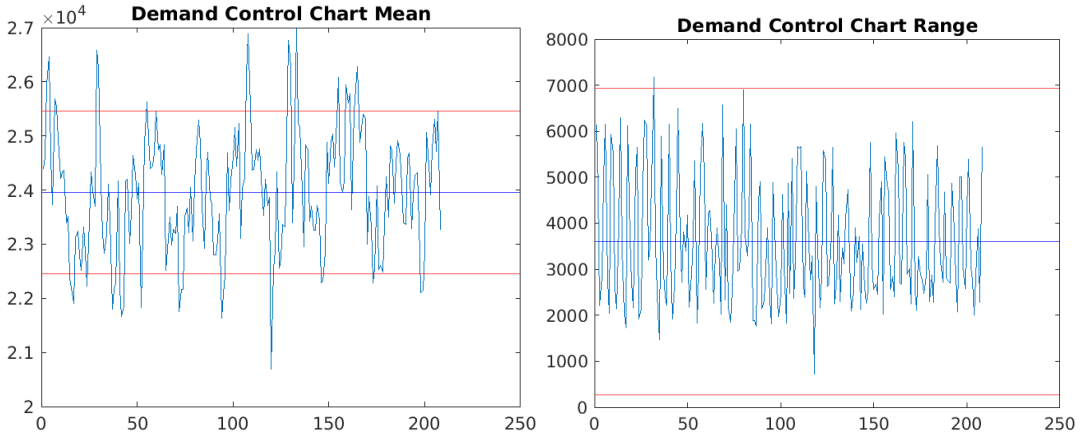


All three variables taken into account show a very similar behaviour of volatility over time and all of them have a usually high volatility this means that even if they are different and a have a different dispersion, it possible for the data of each variable to be dispersed similarly.

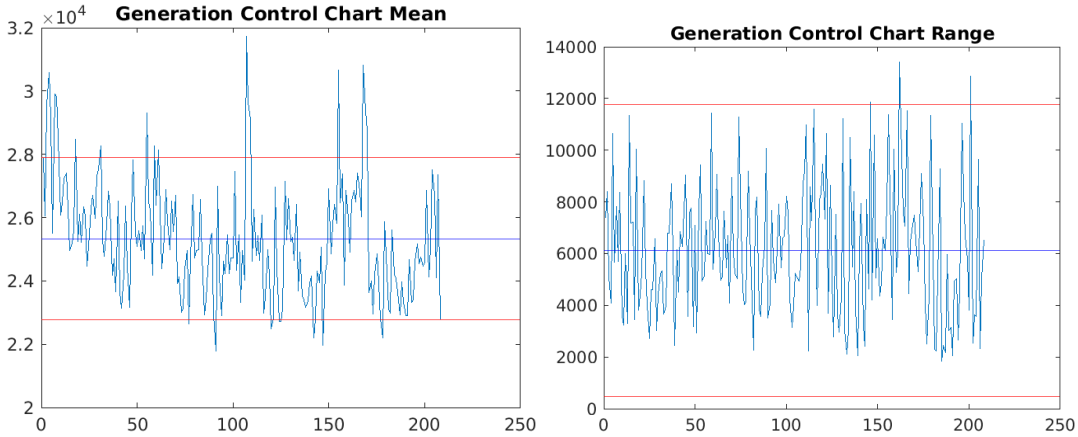
3.3.2 Control Charts

A complete stability analysis is made on an $\bar{x} - R$ chart of the electric energy demand, generation and price over time, samples of 7 hours a week are taken in consideration for the chart, the data1460 is used for this process in each of the three variables taken in count.

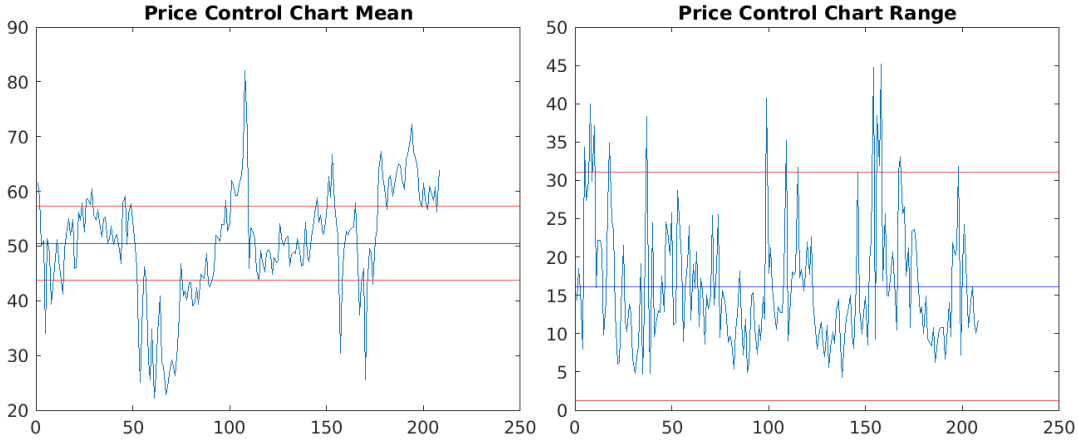
[GD09]



As shown in the previous chart, electric energy demand in Valencia is not under statistic control as there are some outliers with exponential behaviour that surpass the established control limits. The results were predictable as demand has a high correlation with some variables in the dataset but also some other things like special days affect a lot in the quantity of energy demanded.



As shown in the mean control chart, energy generation is a less under control process that has shown a decreasing behaviour due to both, a decrease in resources to produce energy and the increase in the import of electric energy from other cities.



Lastly the price has shown to be the less controlled variable taken in count, this is a predictable outcome due to the low correlation even with the supply and demand so a lot of other unknown variables affect the price and is harder to generate an accurate prediction for the price of energy than to generate a good prediction for the energy demand at a given time.

3.3.3 Exponential smoothing

Exponential smoothing is another way of estimating a data series' volatility when it isn't constant. The method consists on generating a smooth function out of raw data, one that describes the general tendency of data over time. The simplest form of an exponential smoothing algorithm is given by the following formula [NB00]:

$$s_0 = x_0$$

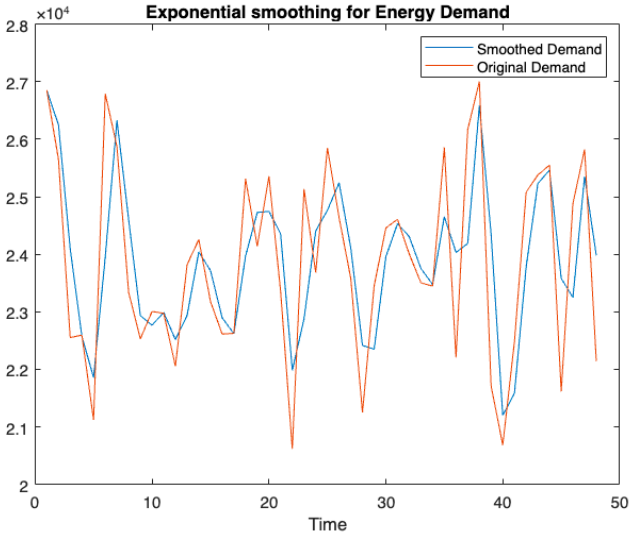
$$s_t = \alpha x_t (1 - \alpha) s_{t-1}, t > 0$$

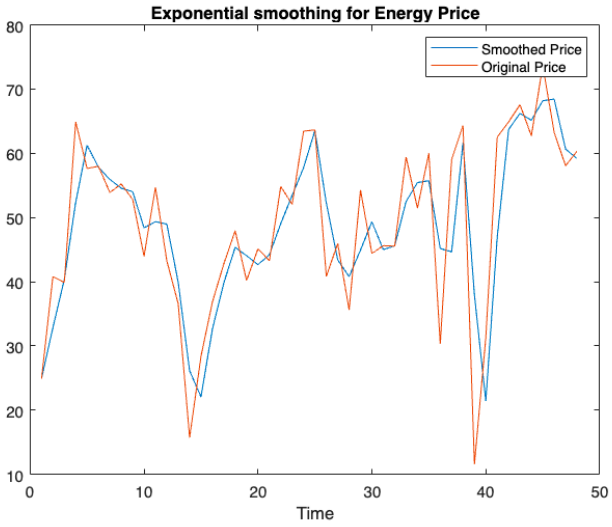
where s_t is the estimation of what the next value of x will be. α is the smoothing factor.

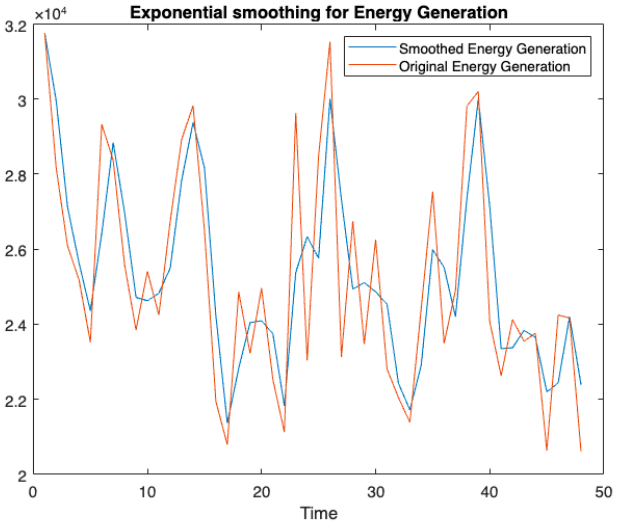
Note that the expression for s_t is a linear combination between the xth observation and and s_{t-1}, meaning that α must be between zero and one. Note that the method works by assigning a higher value for the most recent values. This relation is exponential, thus explaining the name of the method.

[NB00] This is exactly why this method is so efficient to make predictions, a process that will be discussed later on the article.

(Realizar analisis de las figuras)

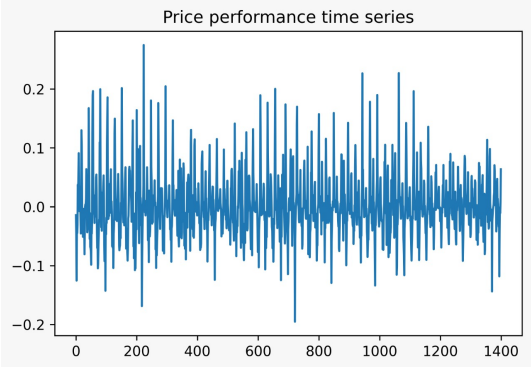


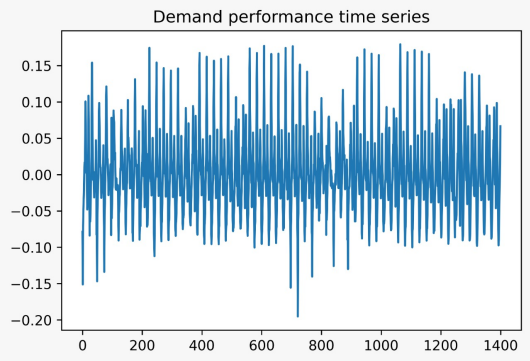




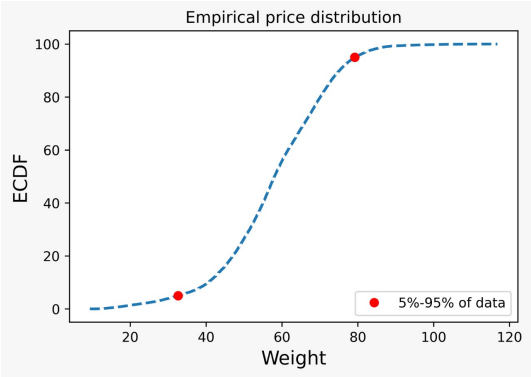
3.4 Performance and prediction error

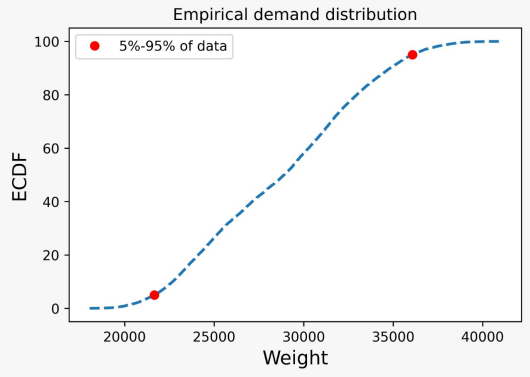
(Explicar que se hizo y que dataset se uso)
(Analisis)





¿Esto lo vamos a incluir?? ¿Que analisis haremos?





3.5 Extreme values and outliers

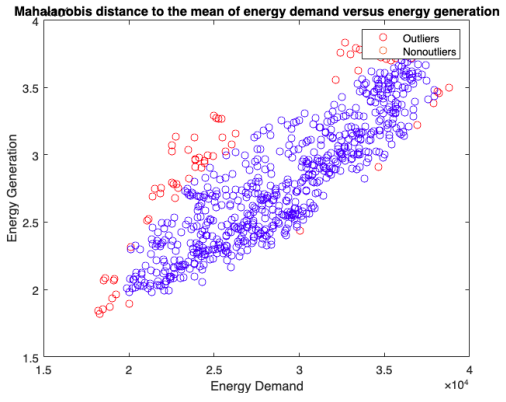
3.5.1 Mahalanobis distance

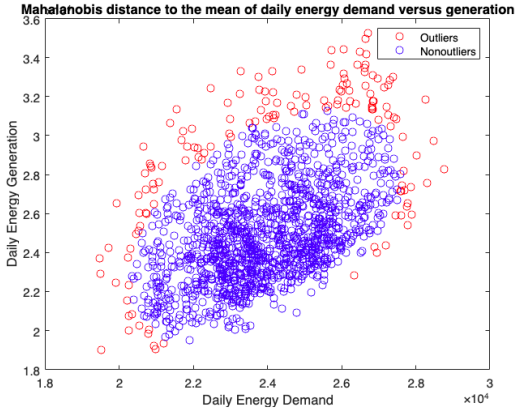
When analyzing data, it is important to detect the presence of outliars, in order to avoid making decisions that would compormise more accurate results.

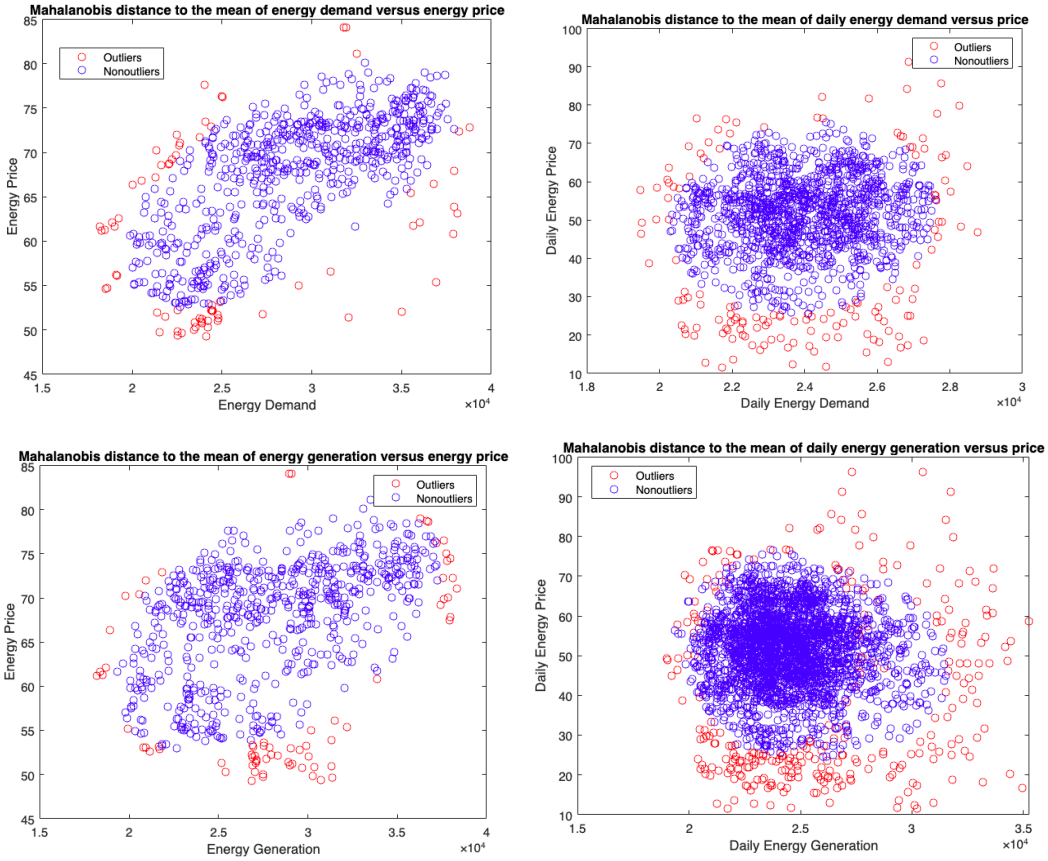
Mahalanobis distance is used to detect outliers in multivariant testing.[Gho19] The distance of an observation $\vec{x} = (x_1, x_2, x_3 ... x_n)^T$, from a set of observations with means $\vec{\mu} = (\mu_1, \mu_2, \mu_3, ..., \mu_n)$ and a given covariance matrix Σ is defined as:

$$D_m(\vec{x} = \sqrt{((\vec{x} - m\vec{u})^T \Sigma^{-1} (\vec{x} - m\vec{u}))}$$

As



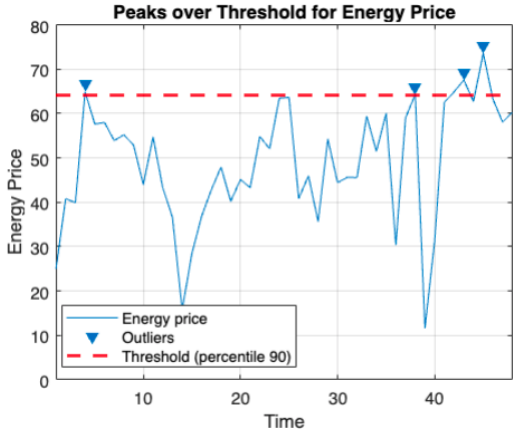


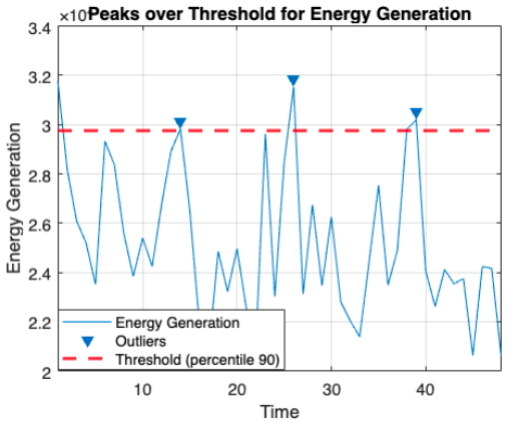


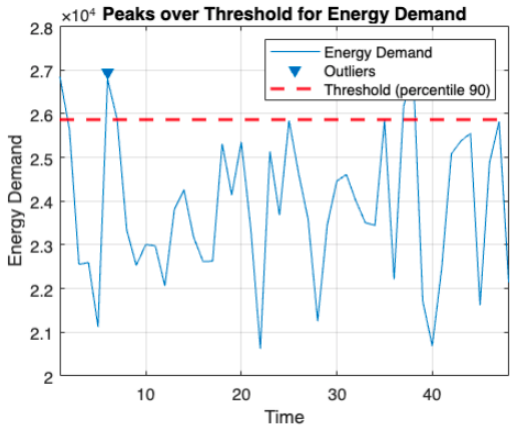
3.5.2 Peaks over Threshold

Peaks over Threshold (POT) identifies the extreme values of the series as those who go over a certain threshold t and these outliers are known as excess returns. In this case, the threshold t es the 90 percentile. The graphs for each of the variables of interest POT's are shown bellow. [MB06]

They show that ...(Analisis)







3.6 Measuring dependency

3.6.1 Autocorrelation Function

Autocorrelation measures the relationship between an observation at time t and its previous observations. It helps detects non randomness in a dataset. Autocorrelation function (ACF) is an specific method for measuring a time series autocorrelation, and measures the correlation between y_t and y_{t+k}, with $k = 0, 1, ..., K$. The difference of time between the previous and current times is referred to as lag k, and its autocorrelation is

$$r_k = \frac{c_k}{c_0}$$

where

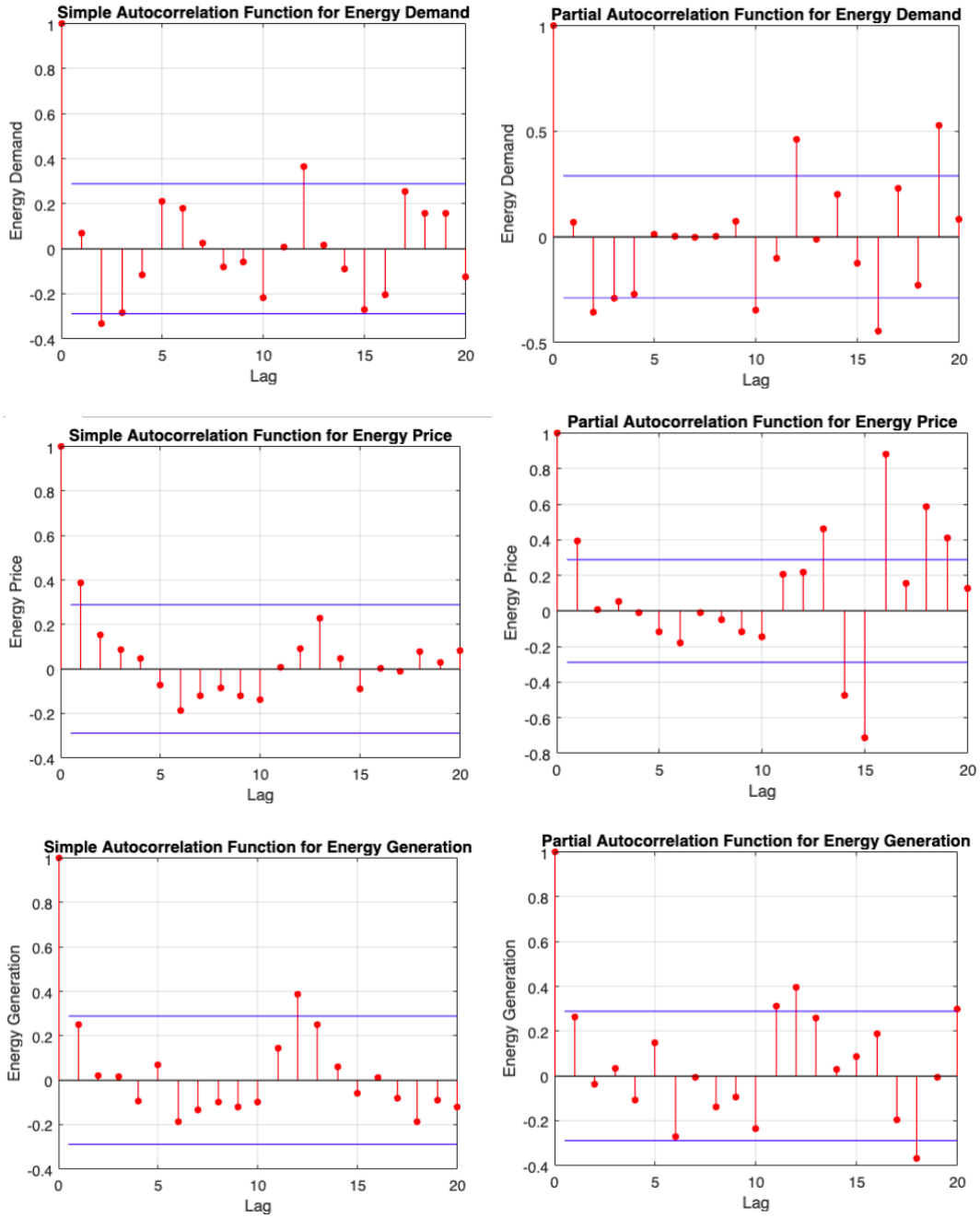
$$ACF(k) = \frac{1}{T}\sum_{t=1}^{T-k}(Y_t - \bar{Y})(Y_{t+k} - \bar{Y})$$

and c_0 is the sample variance.[BJ90]

Partial autocorrelation function PACF is the correlation at lag k is the correlation of values at k time steps apart and all the values in between, different from ACF that only calculates the two values at the intervals' limits.[Dat19]

$$PACF(k) = \frac{cov(Y_i, Y_{i-k}|Y_{i-1}, ..., Y_{i-k+1})}{\sqrt{var(Y_i|Y_{i-1}, ..., Y_{i-k+1}) * var(Y_{i-k}|Y_{i-1}, ..., Y_{i-k+1})}}$$

The ACF and the PACF were estimated for data720, and the results are shown bellow: (Analisis)



3.6.2 Copula

A copula helps detect dependency structures in multivariate data. "A copula is defined as a function that joins multivariate distribution functions to their one-dimensional marginal distribution functions. It is a multivariate distribution function defined on the unit n-cube $[0,1]^n$ ". [Kpa07]

An approach to estimating copulas in a time series with an unknown distribution is to calculate it with the empirical function, which is an empirical measure of the sample and converges with probability 1 to the underlying distribution.

$$\bar{F}_n(t) = \frac{1}{n} \sum_{t=1}^{T-k} IX_i \leq (t)$$

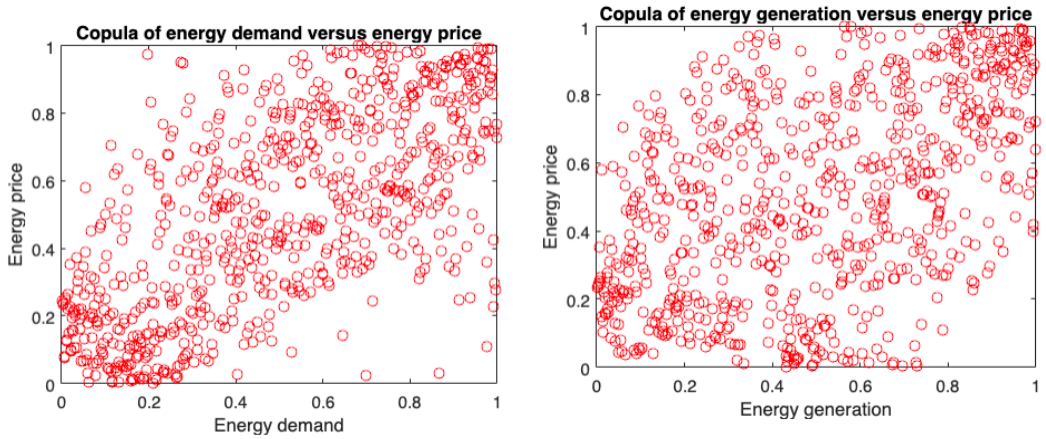
where I is the indicative function.

Then, the bivariate empirical copula function is:

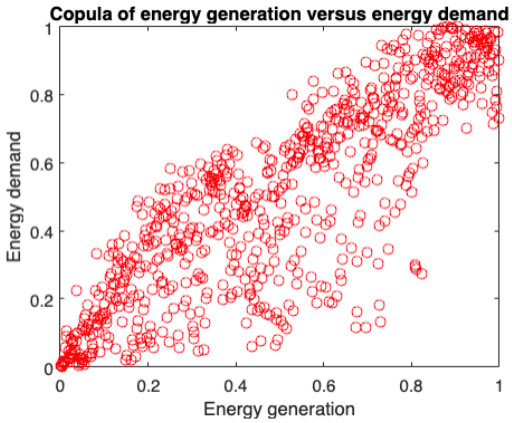
$$C_n(u,v) = \frac{1}{n}\sum_{t=1}^{T-k} I[X_i \leq (t)F_n(u), Y_i \leq (t)G_n(v)]$$

[Kpa07]

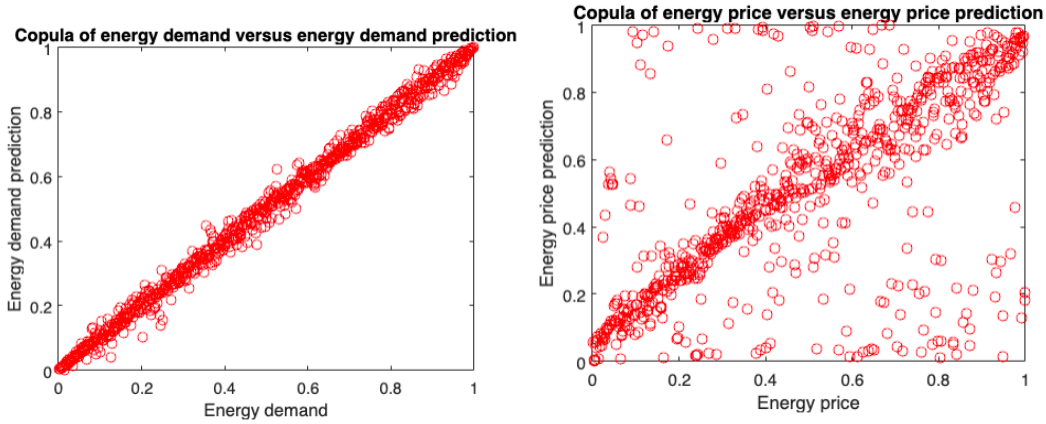
The figures bellow show the copulas found for different pairs of the variables of interest:



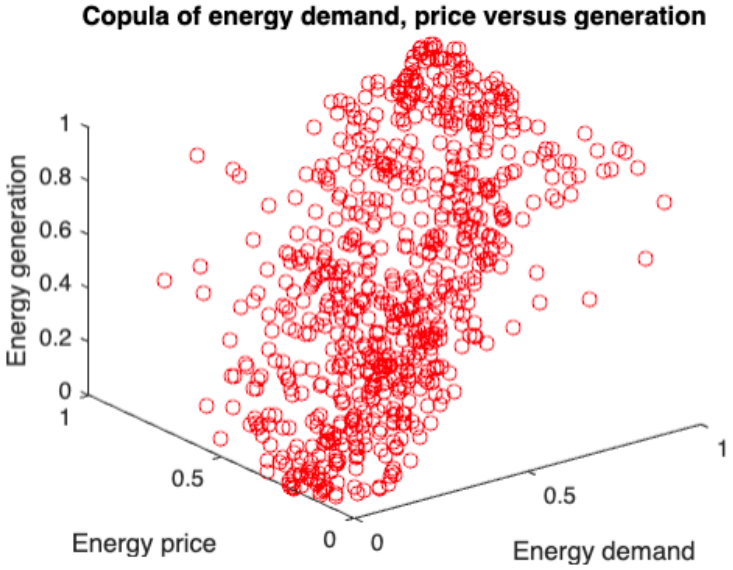
The previous copulas depict a medium structure dependency between the energy demand and the energy generation compared to the energy price in a given time, this means that for making a almost perfect prediction for the price of energy a lot of more different variables are needed.



The copula shows us a more structured relation between the energy generation and energy demand very likely making demand one of the most important variables in creating a prediction for the power generation.



Both this copulas show how good are the predictions made in the data sets compared with the actual values of demand and price respectively, as expected, the predictions on demand had less outliers and a better correlation than the predictions done on the price.



In the three dimensional copula of generation, demand and price is seen a better shaped structure which means there is still dependency.

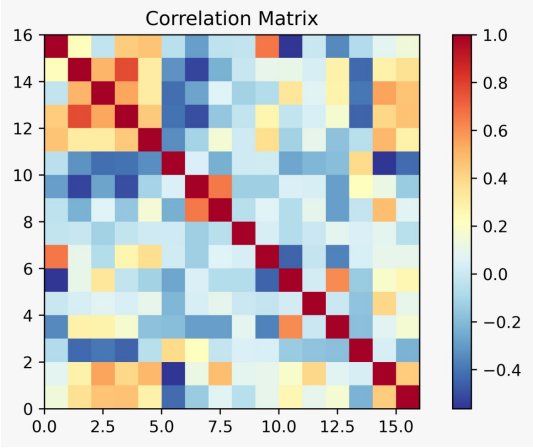
3.6.3 Correlation Coefficient

Correlation is an analysis of two or more different variables which measures their level of relationship. The correlation coefficient R^2 determines how much a variable can be explained in terms of the others, and is measured between $-1 \leq R^2 \leq 1$. The closer the value is to zero, the more independent the variables are from each other and the closer the value is to one, the more dependent the variables are. The sign of R^2 determines the nature of the relationship, positive of negative. (citar con alguna fuente confiable xdxd).

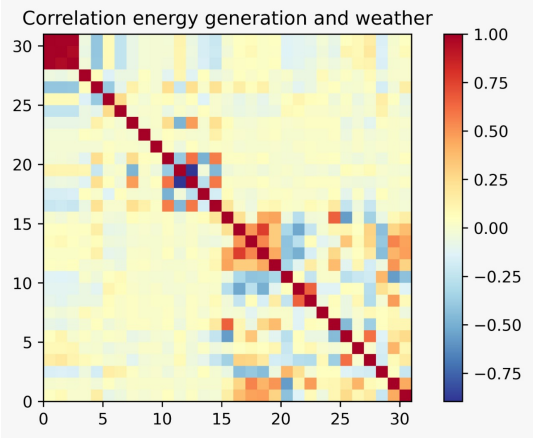
Hacemos otra version robusta?? Spearman? Pearson? Kendall? Dejamos las tablas?? Como ponemos l leyenda??

Coeficiente de correlacion (R2)

Decir que es esto y explicar que se realizo con cov robusta. Explicar leyendas . Camiiii



4 Energy correlation with weather



5 Prediction

Simple Exponential Smoothing for Time Series Forecasting [Van19a] [Van19b]
prediction based on weather??

6 Conclusions

????

7 Future work

?

References

[BJ90] George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Inc., USA, 1990.

[Dat19] Sachin Date. Understanding partial auto-correlation. *Towards Data Science*, 2019.

[GD09] Humberto Gutiérrez Pulido and Román De La Vara Salazar. *Control Estadístico de la Calidad y Seis Sigma*, volume Segunda Edición. McGraw Hill, 2009.

[Gho19] Hamid Ghorbani. Mahalanobis distance and its application for detecting multivariate outliers. *Facta Universitatis Series Mathematics and Informatics*, 34:583, 2019.

[Jha19] Nicolas Jhana. Hourly energy demand generation and weather. *Kaggle*, 2019.

[Kpa07] Tchilabalo Kpanzou. Copulas in statistics. *University of Stellenbosch*, 2007.

[lto21] ltorov. *Hourly-energy-demand-and-prices-an-analysis-on-risk-measuresand-correlation*. Github, 2021.

[MB06] Luis Fernando Melo Velandia and Oscar Becerra Camargo. Medidas de riesgo, características y técnicas de medición: una aplicación del var y el es a la tasa interbancaria de colombia. *Banco de la República, Gerencia Técnica*, 2006.

[NB00] Mohamed N. Nounou and Bhavik R. Bakshi. Chapter 5 - multiscale methods for denoising and compression. In Beata Walczak, editor, *Wavelets in Chemistry*, volume 22 of *Data Handling in Science and Technology*, pages 119–150. Elsevier, 2000.

[Van19a] Nicolas Vandeput. Simple exponential smoothing for time series forecasting. *Towards Data Science*, 2019.

[Van19b] Nicolas Vandeput. Simple exponential smoothing in python from scratch. *Towards Data Science*, 2019.

[Yad19] Dinesh Yadav. Categorical encoding using label-encoding and one-hot-encoder. *Towards Data Science*, 2019.